УДК 519.217.4:81'42

Крамаренко А.А., Некрасов К.А., Филимонов В.В.,
Живодёров А.А., Амиева А.М.

КЛАССИФИКАЦИЯ РУССКОЯЗЫЧНЫХ ТЕКСТОВ ПО СТИЛЯМ НА ОСНОВЕ МОДЕЛИ СЛУЧАЙНЫХ БЛУЖДАНИЙ

Аннотация. В статье предложен формальный подход к исследованию текстов, основанный на модели случайных блужданий. Частотам и взаимному расположению гласных букв текста сопоставлен процесс диффузии квазичастиц. Показано, что параметры диффузии квазичастиц статистически значимо различаются для текстов, соответствующих различным функциональным стилям языка, так что могут быть использованы для классификации текстов. Выделено пять групп текстов, различимых рассмотренным методом.

Ключевые слова: модель случайных блужданий, функциональные стили языка, анализ текстов.

Abstract. A formal approach to text analysis is suggested that is based on the random walk model. The frequencies and reciprocal positions of the vowel letters are matched up by a process of quasi-particle migration. Statistically significant difference in the migration parameters for the texts of different functional styles is found. Thus, a possibility of classification of texts using the suggested method is demonstrated. Five groups of the texts are singled out that can be distinguished from one another by the parameters of the quasi-particle migration process.

Keywords: the random walk model, the language functional styles, text analysis.

Введение

В настоящее время всё больше растёт интерес к формальному анализу текстов. Так, впервые математические методы для анализа текстов использовал А.А. Марков [1]. Впоследствии было развито определение авторства текста с помощью цепей Маркова [2]. Кроме того, широко применяется и модифицируется закон Ципфа для изучения естественных языков на основе массивов текстов. К примеру, в работе [3] рассматривается несколько моделей, описывающих частотные распределения всех букв алфавита для разных языков.

Одним из перспективных направлений также является кластеризация текстов по каким-либо критериям. Среди современных авторов в этой области работают Ю.Н. Орлов и К.П. Осминин. Ими были рассмотрены распределения отдельных букв и пар букв, а также с помощью выборочных функций распределения проведена кластеризация художественных текстов по жанрам и авторам [4].

В рамках данной работы решается проблема автоматизированной классификации русскоязычных текстов по функциональным стилям речи на

основе анализа частот гласных букв и их взаимного расположения. Такой выбор критерия классификации позволяет охватить множество текстов от художественных до законотворческих.

Необходимо отметить, что использование математических методов исследования текстов позволяет опираться только на некие внутренние структуры текста, не учитывая смысловую составляющую. Это даст возможность проводить автоматический анализ текстов без участия экспертов.

Теоретическая часть

Метод анализа текстов, представленный в данной работе, основан на аналогии между тепловым движением частиц и перемещением неких квазичастиц по траекториям, которые задаются структурой текста. По закону Эйнштейна средний квадрат смещения частиц в отсутствие действия внешних сил прямо пропорционален времени (1):

$$<R^2> = 2nDt, \quad (1)$$

где $<R^2>$ – средний квадрат смещения частицы; D – коэффициент диффузии; t – время. Рассматривается двумерный случай: $n = 2$.

Аналогия между законом Эйнштейна (1) и процессом перемещения квазичастиц состоит в том, что переход к следующей гласной букве эквивалентен переходу к следующему моменту времени. В качестве момента времени рассматривается порядковый номер буквы от начала текста без учёта согласных букв.

Текст задаёт перемещение квазичастицы следующим образом. На первом шаге квазичастица имеет нулевые координаты. Очередная гласная буква из текста представляет некое правило перехода и задаёт вектор смещения частицы (Рисунок 1).

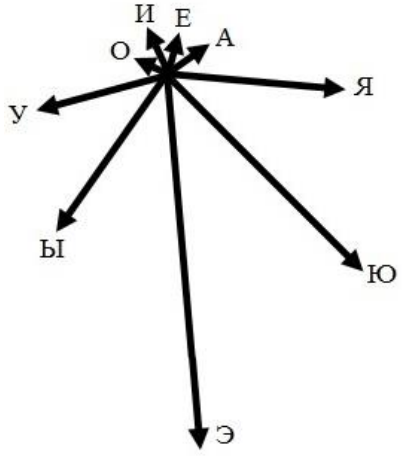

Рисунок 1 – Примерная диаграмма векторов

На Рисунке 1 строго изображены направления векторов. Соотношение длин векторов соответствует реальному только качественно, из-за большой разницы длин, действительные значения для конкретного текста («Теоретическая физика», том 2, Ландау Л.Д., Лифшиц Е.М.) приведены в Таблице 1.

Таблица 1 – Параметры правил перехода

Буква	ω	|r|	α
А	0,162	6,168	35
Е	0,222	4,500	75
И	0,185	5,414	115
О	0,251	3,992	155
У	0,051	19,533	195
Ы	0,042	23,764	235
Э	0,013	79,763	275
Ю	0,017	58,843	315
Я	0,058	17,350	355

Модуль вектора смещения обратно пропорционален частоте ω появления буквы в тексте, чтобы исключить дрейф в сторону с большими частотами букв. Соответствие угла букве задано произвольно, шаг между углами равен 360°/9 = 40°, так как мы рассматриваем 9 гласных букв. На Рисунке 2 показаны траектории первых трёх квазичастиц для текста «Теоретическая физика», том 2, Ландау Л.Д., Лифшиц Е.М.

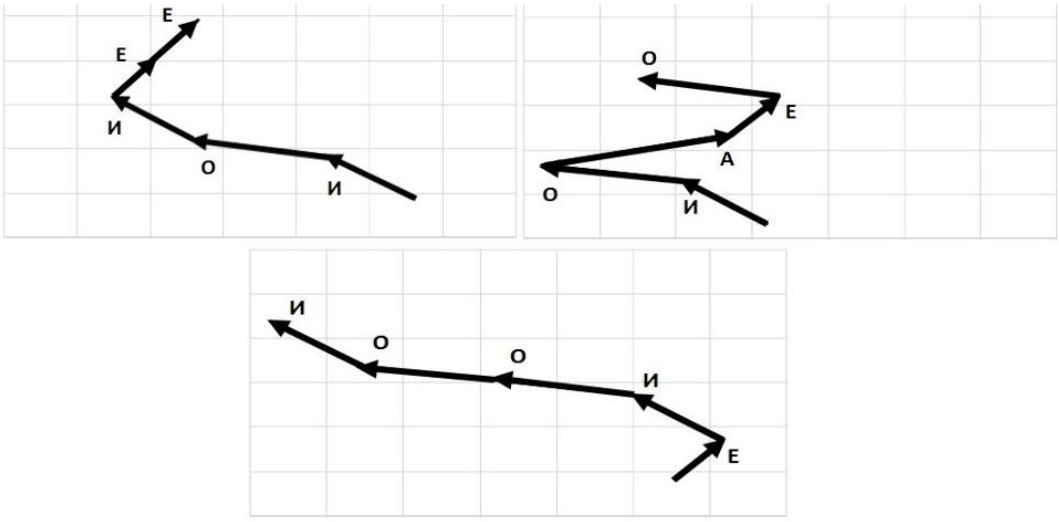

Рисунок 2 – Примеры траекторий квазичастиц для текста «Теоретическая физика», том 2, Ландау Л.Д., Лифшиц Е.М.

Перемещение квазичастицы вычислялось как сумма вектора текущих координат квазичастицы и смещения, задаваемого очередной буквой. Усреднение значений $R^2(t)$ проводилось по совокупности квазичастиц, количество которых определялось длиной текста. На каждую квазичастицу в настоящей работе приходилось 100 гласных букв, задававших 100 шагов перемещения.

Для анализа групп текстов был введён коэффициент *D*, равный тангенсу угла наклона прямой $<R^2(t)>/(2n)$. Он определялся методом наименьших квадратов.

Для некоторых текстов наблюдалось очевидное отклонение зависимости $<R^2(t)>$ от линейного закона. Для описания степени отклонения введена относительная поправка к закону Эйнштейна (RC), которую определяли следующим образом.

Пусть зависимость среднего квадрата смещения квазичастицы от времени описывается функцией *f(x)*. Разложим эту функцию в ряд Тейлора:

$$f(x) = f_0 + f' \cdot x + f'' \cdot \frac{x^2}{2} + \ldots \qquad (2)$$

Будем рассматривать только прямую $f_1(x)$ и полином второй степени $f_2(x)$. Пусть $f(x) \approx f_2(x)$, поскольку полином второй степени точнее прямой. Тогда

$$\Delta f = f(x) - f_1(x) = f_2(x) - f_1(x) = f_0 + f' \cdot x + f'' \cdot \frac{x^2}{2} - f_0 - f' \cdot x = f'' \frac{x^2}{2} \qquad (3)$$

Если рассмотреть отличие реального процесса от диффузионного, где диффузионный процесс представляет прямая, то

$$RC = \frac{\Delta f(x)}{f_1(x)} = \frac{f'' \cdot \frac{x^2}{2}}{f_0 + f' \cdot x} = \left\| f_0 \approx 0 \right\| = \frac{f'' \cdot \frac{x^2}{2}}{f' \cdot x} = \left. \frac{a_2 \cdot x}{a_1} \right|_{x=t} = \frac{a_2 \cdot t}{a_1}, \qquad (4)$$

где a_2 – коэффициент при старшем члене полинома второй степени, a_1 – коэффициент при старшем члене прямой.

Объект и метод исследования

Объект исследования – специально созданный Корпус текстов русского языка. На сегодняшний день он включает в себя около 1500 текстов и постоянно пополняется. Каждый подкорпус представляет собой тот или иной функциональный стиль речи: поэзия (poet), художественная проза (pr), научный (sc), официально-деловой (off), публицистический (pub) и религиозный (rel). Произведения зарубежных писателей и религиозные тексты представлены в переводе на современный русский язык.

Исследования проводились только на гласных буквах, причём буквы е и ё отождествлялись. Из каждого подкорпуса было взято по 20 текстов с количеством гласных букв от 100 до 500 тысяч для достоверности расчётов. Для некоторых подкорпусов (поэзия – 13, публицистика – 16) текстов было взято меньше, поскольку трудно найти тексты одного автора необходимой длины.

Каждый текст был разделён на фрагменты по 100 гласных букв. Фрагмент соответствует квазичастице, соответственно, количество квазичастиц находилось в диапазоне от 1 до 5 тысяч. Проведено усреднение по этим фрагментам, построена зависимость среднего квадрата смещения от времени и проведена аппроксимация зависимости полиномами первой и второй степени.

Результаты

В результате расчётов получены траектории движения квазичастиц, соответствующие рассмотренным текстам, для каждого из них определены значения коэффициента D и относительной поправки RC, а также средние значения этих параметров по подкорпусам (Таблица 2).

Таблица 2 – Средние значения параметров текстов по подкорпусам

Подкорпус	D	σ^2 [D]	RC, %
poet	103,20±16,95	690,51	14,83±5,05
pr	85,35±7,21	256,89	10,84±3,20
sc	77,74±6,85	232,04	18,57±3,52
off	175,73±29,46	4302,83	37,94±8,41
pub	76,23±3,57	62,17	9,92±3,53
rel	113,92±18,23	1644,04	15,66±3,10

На Рисунках 3 и 4 представлены выборки значений D и RC для текстов по каждому подкорпусу. Значения коэффициента D распределились по двум диапазонам: $D_1 < 125$; $D_2 > 125$. В первом диапазоне находятся тексты из подкорпусов поэзии, художественной прозы, научные тексты и публицистика. Во втором – административные тексты. Религиозные тексты расположились на границе между этими диапазонами.

Диапазоны значения RC оказались следующими: $RC_1 < 10\%$; $10\% < RC_2 < 25\%$; $RC_3 > 25\%$. В первый интервал попали тексты художественной прозы и публицистики, во второй – научные и религиозные тексты, в третий – административные. Поэзия оказалась на стыке между первым и вторым диапазонами.

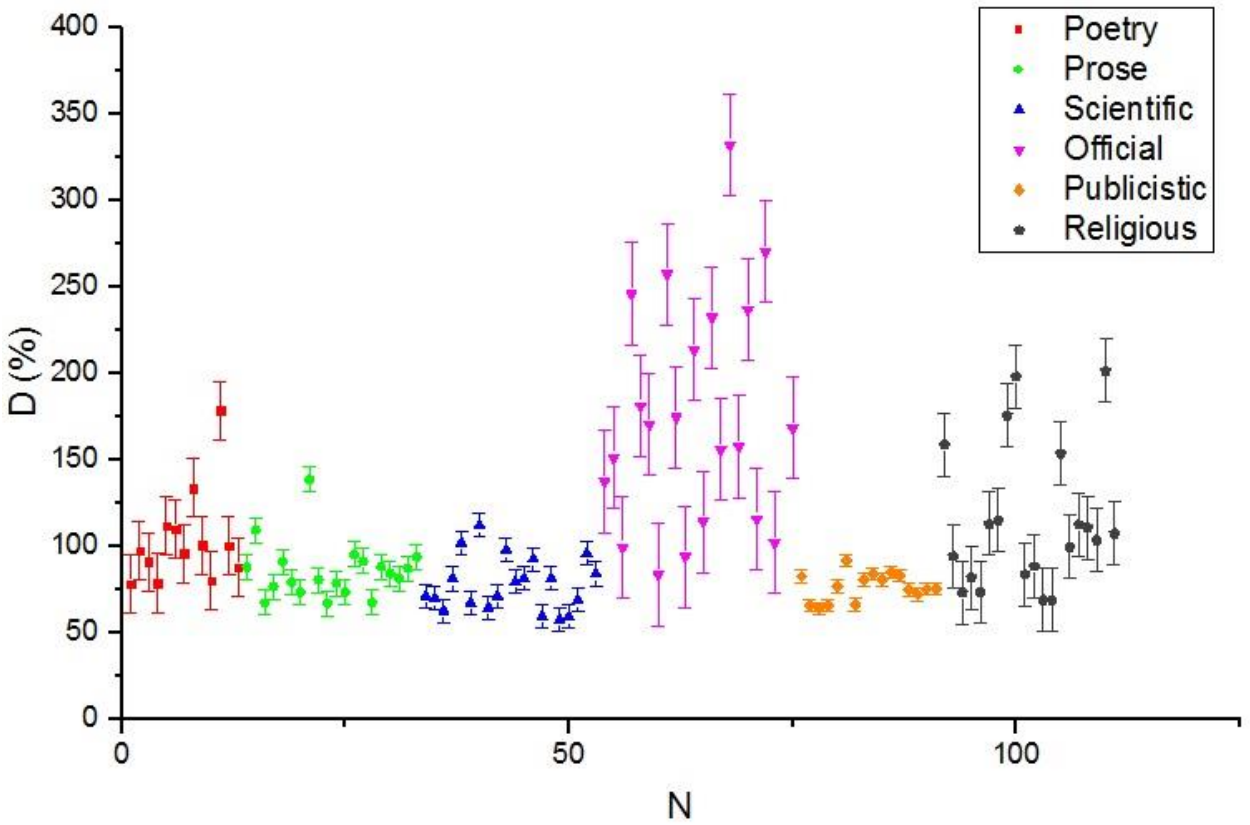

Рисунок 3 – Значения D проанализированных текстов
N – номер текста

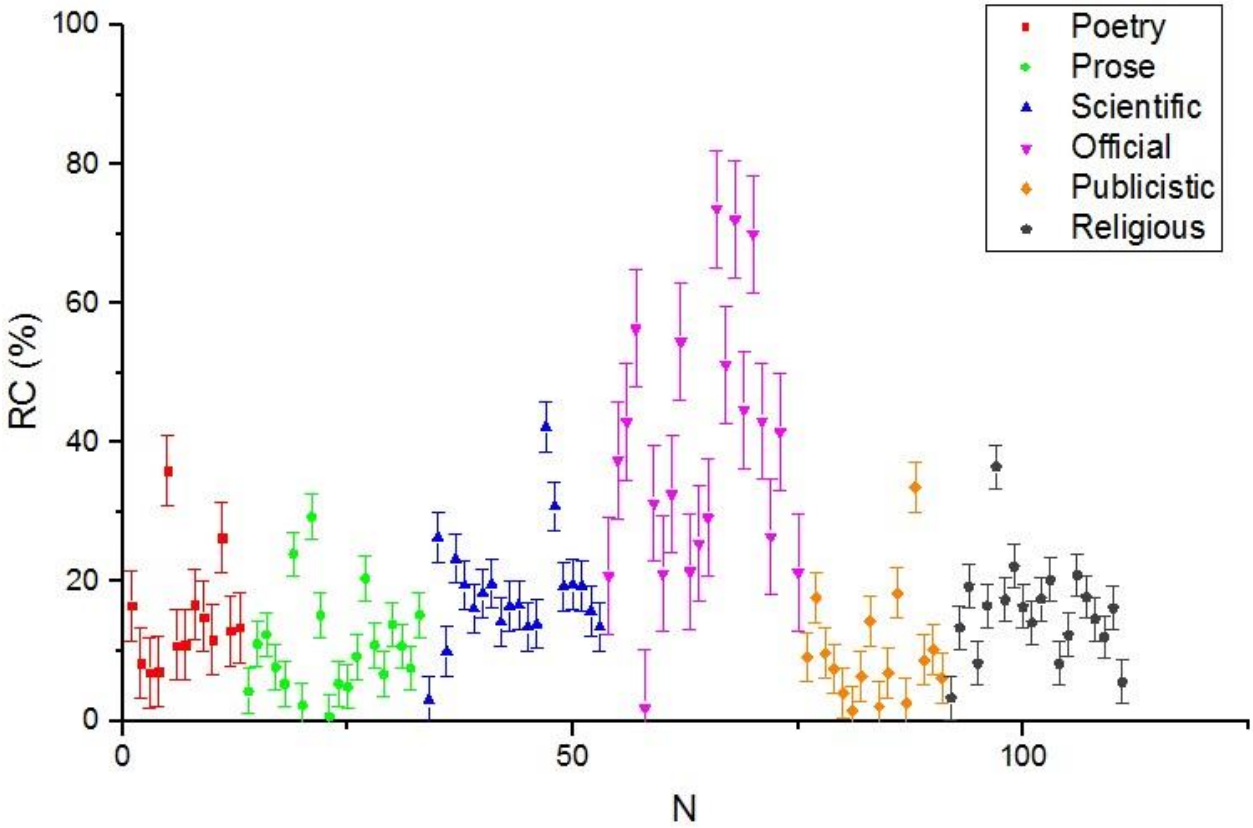

Рисунок 4 – Значения RC проанализированных текстов
N – номер текста

Для количественной оценки различения групп был использован критерий χ^2. Рассмотрены два случая. В первом случае оценивалась статистическая значимость отличия каждой из групп от генеральной совокупности текстов. Гипотеза H_0 для первого случая: распределения выборок одинаковы и являются нормальными.

$$\chi^2 = n \cdot \sum_{i=1}^{k} \frac{(\frac{n_i}{n} - P_i)^2}{P_i}, \quad (5)$$

где n – количество всех текстов выборки, n_i – количество текстов, попавших в i-й интервал, P_i – теоретическая вероятность попадания всей совокупности текстов в i-й интервал, k – количество интервалов.

Диапазон значений D был разделён на 7 интервалов, определена теоретическая и экспериментальная вероятность попадания значения D в каждый из интервалов. Для простоты теоретические распределения значений D были приняты нормальными.

В результате для каждого подкорпуса были получены значения критерия χ^2 (Таблица 3).

Таблица 3 – Значения критерия χ^2 для первого случая

Подкорпус	χ^2
poet	9,102
pr	15,032
sc	15,020
off	16,528
pub	11,762
rel	14,952

Было проведено сравнение полученных значений критерия χ^2 с критическим значением критерия χ^2 для данных степеней свободы r = k – m – 1 = 4, где k = 7 – число интервалов, m = 2 – количество параметров нормального распределения. Выяснилось, что на уровне значимости $\alpha = 0{,}06$ отличия статистически значимы.

Во втором случае оценивалась статистическая значимость отличия каждой группы от других попарно. Гипотеза H_0 для второго случая: распределение каждой выборки не отличается от любой другой выборки и является нормальным. Значения критерия χ^2 рассчитывались по формуле:

$$\chi^2 = n_j \cdot \sum_{i=1}^{k} \frac{(\frac{n_{ij}}{n_j} - P_{ig})^2}{P_{ig}}, \quad (6)$$

где n_j – количество всех текстов выборки j, n_{ij} – количество текстов выборки j, попавших в i-й интервал, P_{ig} – теоретическая вероятность попадания текстов в i-й интервал для g-й выборки, k – количество интервалов.

Каждая выборка по очереди сравнивалась с остальными по схеме, приведённой выше. В результате для каждого подкорпуса были получены значения критерия χ^2 (Таблица 4).

Таблица 4 – Значения критерия χ^2 для второго случая

Подкорпус	poet	pr	sc	off	pub	rel
poet		17,38	17,41	$7{,}5 \cdot 10^8$	13,63	17,73
pr	396,70		17,40	$3{,}9 \cdot 10^4$	13,56	$6{,}1 \cdot 10^5$
sc	$1{,}9 \cdot 10^4$	16,56		$5{,}0 \cdot 10^6$	12,85	$1{,}3 \cdot 10^8$
off	10,77	17,67	17,62		14,17	17,34
pub	$1{,}5 \cdot 10^4$	$4{,}3 \cdot 10^3$	18,11	$1{,}2 \cdot 10^{11}$		20,53
rel	10,00	16,34	16,36	46,66	12,81	

При сравнении значений критерия χ^2 с критическим значением критерия χ^2 для r = 4 выяснилось, что на уровне значимости $\alpha = 0{,}05$ отличия статистически значимы.

Обсуждение

Одновременное рассмотрение параметров D и RC показывает, что тексты распадаются на пять групп в соответствии с выделенными диапазонами (Таблица 5).

Таблица 5 – Критерий χ^2 для второго случая

Подкорпус	D	RC
poet	D_1	RC_1-RC_2
pr+pub	D_1	RC_1
sc	D_1	RC_2
off	D_2	RC_3
rel	D_1-D_2	RC_2

Таким образом, получено пять групп с различными параметрами.

Каждая выборка по подкорпусу статистически значимо отличается от всей совокупности рассмотренных текстов. Полученные выборки с высокой степенью значимости отличаются друг от друга. Следовательно, предложенный метод классификации текстов является перспективным.

Библиографический список

1. Марков А. А. Примѣръ статистическаго изслѣдованія надъ текстомъ «Евгенія Онѣгина», иллюстрирующій связь испытаній въ цѣпь / А. А. Марков : (доложено в заседании физ.-мат. отд-ния 23 янв. 1913 г.) // Санкт-Петербург : Тип. Имп. Акад. наук, 1913. – С. 153–162. – Отт. из: Известия Императорской Академии Наук. – 1913.
2. Хмелёв Д. В. Распознавание автора текста с использованием цепей А. А. Маркова / Д. В. Хмелёв // Вестник МГУ. Сер. 9: Филология. – 2000. – № 2. – С. 115–126.
3. Гельфанд М. С. О ранговых распределениях частот букв в естественных языках / М. С. Гельфанд, Чжао Минь // Проблемы передачи информации. – 1996. – Т. 32, вып. 2. – С. 89–95.
4. Орлов Ю. Н. Определение жанра и автора литературного произведения статистическими методами / Ю. Н. Орлов, К. П. Осминин // Прикладная информатика. – 2010. – № 2 (26). – С. 95–108.